The Prodromal Questionnaire: a case for IRT-based adaptive testing of psychotic experiences?

van Bebber, J.; Wigman, J.T.W.; Meijer, R.R.; Ising, H.K.; van den Berg, D.P.G.; Rietdijk, J.; Dragt, S.; Klaassen, R.M.C.; Nieman, D.H.; de Jonge, P.; Sytema, S.; Wichers, M.; Linszen, D.H.; van der Gaag, M.; Wunderink, L.

published in
International Journal of Methods in Psychiatric Research
2017

DOI (link to publisher)
10.1002/mpr.1518

document version
Publisher's PDF, also known as Version of record

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. (2016)
Published online in Wiley Online Library
(wileyonlinelibrary.com) **DOI:** 10.1002/mpr.1518

The Prodromal Questionnaire: a case for IRT-based adaptive testing of psychotic experiences?

JAN VAN BEBBER,[1,2] JOHANNA T.W. WIGMAN,[1,2,3] ROB R. MEIJER,[4] HELGA K. ISING,[5] DAVID VAN DEN BERG,[5] JUDITH RIETDIJK,[5] SARA DRAGT,[6] RIANNE KLAASSEN,[7] DORIEN NIEMAN,[6] PETER DE JONGE,[1] SJOERD SYTEMA,[1] MARIEKE WICHERS,[1] DON LINSZEN,[6] MARK VAN DER GAAG[5,8] & LEX WUNDERINK[2]

1 Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University Medical Center Groningen, The Netherlands
2 Department of Education and Research, GGZ Friesland, Leeuwarden, The Netherlands
3 University Medical Center Groningen, Rob Giel Research Center (RGOc), University of Groningen, The Netherlands
4 Department of Psychometrics and Statistics, University of Groningen, Groningen, The Netherlands
5 Department of Psychiatry, Parnassia Psychiatric Institute, The Hague, The Netherlands
6 Department of Psychiatry, Academic Medical Center, Amsterdam, the Netherlands
7 Department of Child and Adolescent Psychiatry, University Medical Center Utrecht
8 VU University and EMGO+ Institute for Health and Care Research, Amsterdam, The Netherlands

Key words
psychosis, item response theory, computerized adaptive tests, real data simulations

Correspondence
Jan Van Bebber, University Medical Center Groningen, Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), PO-box 30001, 9700 RB Groningen, The Netherlands. Telephone (+31) 642506052 Fax (+31) 503619722
Email: j.van.bebber@umcg.nl

Abstract

Computerized adaptive tests (CATs) for positive and negative psychotic experiences were developed and tested in $N = 5705$ help-seeking, non-psychotic young individuals. Instead of presenting all items, CATs choose a varying number of different items during test administration depending on respondents' previous answers, reducing the average number of items while still obtaining accurate person estimates.

We assessed the appropriateness of two-parameter logistic models to positive and negative symptoms of the Prodromal Questionnaire (PQ), computed measurement precision of all items and resulting adaptive tests along psychotic dimensions by Real Data Simulations (RDS), and computed indices for criterion and predictive validities of the CATs. For all items, mean absolute differences between observed and expected response probabilities were smaller than 0.02. CAT-POS predicted transition to psychosis and duration of hospitalization in individuals at-risk for psychosis, and CAT-NEG was suggestively related to later functioning. Regarding psychosis risk classifications of help-seeking individuals, CAT-POS performed less than the PQ-16.

Adaptive testing based on self-reported positive and negative symptoms in individuals at-risk for psychosis is a feasible method to select patients for further

Received 14 January 2016;
revised 2 May 2016;
accepted 3 June 2016

risk classification. These promising findings need to be replicated prospectively in a non-selective sample that also includes non-at-risk individuals.

Introduction

To enable timely intervention in psychosis, its early detection is important (McGorry *et al.*, 2003, 2008). Therefore, there is a great need for efficient and effective screening tools for early expressions of psychosis that can be implemented easily at entry into the medical care system. This study investigated the psychometric properties of the Dutch version of the Prodromal Questionnaire (PQ-92) (Loewy *et al.*, 2005), a screening instrument for psychosis, in order to explore the possibility of building computerized adaptive tests (CATs). Adaptive tests are appealing, because they are short and a large number of domains of psychopathology may be assessed without the need to administer hundreds of items.

At risk for psychosis

There is increasing evidence supporting a continuous view on psychosis (Van Os *et al.*, 1999, 2000, 2009; Johns and Van Os, 2001; Hanssen, 2004; Wigman, 2011). This continuum of psychotic severity ranges from normality through schizotypy to full blown clinical psychotic disorder. Much research focused on the period before onset of a first psychotic episode, called the ultra-high-risk (UHR) period. Individuals at UHR for developing psychosis are defined by the at-risk mental state (ARMS) (Yung and McGorry, 1996; Yung *et al.*, 1998, 2005) criteria: (i) attenuated positive symptoms (APS group), (ii) brief limited intermittent psychotic states (BLIPS group), or (iii) familial liability for psychosis, defined as either having a first degree relative with any psychotic disorder or having a diagnosis of schizotypy (genetic risk group). In addition, individuals must either report persistently low levels of functioning or a recent substantial decline in functioning (van der Gaag *et al.*, 2012) to meet ARMS criteria (McGorry *et al.*, 2003). Adequate recognition of ARMS enables clinicians to offer specific treatment such as cognitive behavioral therapy (van der Gaag *et al.*, 2012) as soon as possible, thereby delaying or even preventing the onset of a first psychotic episode. Furthermore, recognizing individuals at UHR may substantially shorten the duration of untreated psychosis (DUP) should these individuals transition to psychosis. DUP refers to the period between manifestation of the first psychotic symptoms and initiation of adequate treatment(Marshall *et al.*, 2005), and shorter DUP is associated with better prognosis (Wunderink *et al.*, 2009; Chang *et al.*, 2012a, 2012b, 2013). In order to detect ARMS as soon and accurately as possible, one strategy is to first screen patients with self-report inventories of psychotic symptoms and then, if they score above a cutoff, assess semi-structured interviews that tap the same symptom dimensions more in-depth. This two-stage strategy increases the sensitivity and specificity of diagnostic classifications, differentiating well between individuals who do or do not develop psychosis according to diagnosis by psychiatrists (Miller *et al.*, 2002; Yung *et al.*, 2003; Loewy *et al.*, 2005).

The Prodromal Questionnaire (PQ)

The PQ-92 is a self-report inventory to be used in this first stage. PQ-92 items are clustered into four domains: positive symptoms (45 items), negative symptoms (19 items), disorganized symptoms (13 items), and general symptoms (15 items). In this paper, we focus on the positive (PQ-92-POS) and negative (PQ-92-NEG) symptom dimensions. Positive symptoms are highly predictive (Loewy *et al.*, 2005; Ising *et al.*, 2012) of the differentiation between healthy and ARMS/psychosis as assessed by structured interviews (Miller *et al.*, 2002; Yung *et al.*, 2005). Negative symptoms are predictive of later social and vocational functioning (Pogue-Geile and Zubin, 1987; Lin *et al.*, 2011). The PQ-16 is a shortened version of the original questionnaire (Ising *et al.*, 2012) that was specifically designed to discriminate optimally between normal and ARMS/psychosis mental states according to the comprehensive assessment of at-risk mental state (CAARMS). It contains those 16 items of the PQ-92 that best predict this differentiation, and the sensitivity and specificity are both 87%.

Computerized adaptive testing (CAT) and item response theory (IRT)

The aim of CAT is to obtain the same measurement precision using fewer items than the original instrument (Wainer, 2010). In clinical applications of CAT, the intensity level of the items to be administered is tailored to the estimated levels of psychopathological symptom

experiences of respondents. That is, in case of dichotomous items, the objective of the algorithm is to present items for which respondents have a chance of approximately 50% of endorsing the item. An advantage of CAT in measuring psychosis is that mainly symptoms are selected that match a patient's severity level, resulting in short questionnaires. Item selection is an iterative process: with each symptom administered, an improved estimate of an individual's symptom severity level is obtained and the next symptom to be administered is the one that yields the most information regarding this individual estimate. This process continues until a certain stop criterion is reached, usually a predetermined level of accuracy, expressed as a maximum tolerable standard error (SE) for the purpose of testing. Further illustration of the principle of adaptive testing is given in the Supportive Information. Adaptive testing is usually based on item response theory (IRT) (Reise and Waller, 2009; Embretson and Reise, 2013), a family of probabilistic models. An IRT model specifies how both respondent's level of symptom severity and item properties influence the response pattern. If the postulated model fits the observed data reasonably well, individual scores are still comparable (may be placed on the same metric), although each respondent gets his/her own set of symptoms that is tailored to their estimated symptom severity levels. Thus, with CAT, each tested person will complete a different set of questions, depending on the number of questions needed to reach a preset threshold of accuracy. An illustration of CAT can be found in the Supportive Information.

Aims of this study

The first aim of this study was to determine whether the positive and negative symptom dimensions of the PQ-92 could be adequately represented by IRT models. The second aim was to assess how many symptoms of each dimension are needed to reach adequate levels of measurement precision. The third aim was to investigate how well the CAT-POS and CAT-NEG predict clinical and functional outcome regardless of ARMS. In order to achieve the second and third aims, we utilized the principle of real data simulations (RDS; see Methods section).

Methods

Data collection design

The data of three interdependent samples gathered in the Dutch Early Detection and Intervention Evaluation (EDIE-NL) trial (van der Gaag *et al.*, 2012) and the variety of instruments and measures that have been used, are presented in Table 1.

1 *Help-seeking*. Help-seeking individuals (N = 5705) were screened with the PQ-92 between February 2008 and February 2010 at four different sites in the Netherlands: $N = 3666$ patients at the Mental Health Center PsyQ Haaglanden, The Hague; $N = 1109$ patients at the Friesland Mental Health Services; $N = 326$ patients at the Mental Health Center Rivierduinen, Leiden and surrounding areas; $N = 276$ at the Mental Health Center PsyQ, Amsterdam; $N = 206$ at the ABC (Altrecht),

Table 1. Flowchart data collection design

	Sample		
	1. Help-seeking	2. UHR + sub-threshold levels of positive symptoms	3. UHR follow-up
Criteria	---	PQ-92-POS > 17 (N = 420) + 11 < PQ-92-POS < 18 (N = 147)	CAARMS POS DSM axis-one/two
N	5705	567	90
Instruments measures	PQ-92	CAARMS, SOFAS	Diagnosis, Hospitalization, SOFAS
POS-PQ	Mean = 11.51, SD = 8.38	Mean = 21.44, SD = 6.27	Mean = 24.07, SD = 6.52
NEG-PQ	Mean = 7.90, SD = 4.94	Mean = 11.67, SD = 3.92	Mean = 12.71, SD = 4.28
Research questions	Model fit Local dependence Properties CATs	Criterion validity	Predictive validity

Utrecht; $N = 116$ patients at the Academic Medical Center, Amsterdam. Six of these individuals were removed from the analyses because they had missing data on all positive symptoms. With respect to positive symptoms, 2.15% of the total data were missing and "I believe in telepathy, psychic forces, or fortune-telling" had the highest percentage of missing values (4.65%). With respect to negative symptoms, 3.08% of the total data were missing and "People find me aloof and distant" had the highest percentage of missing values (4.73%). Mean age was 24.7 (standard deviation [SD] 5.7, range 10–37), and 36.6% were male (63.2% female, .2% missing).

2 *UHR*. A subgroup ($N = 567$) of the first sample was assessed with the CAARMS and the fourth version of the Social and Occupational Functioning Scale (SOFAS) after the intake (see later for instrument descriptions). This subsample included all individuals that endorsed 18 or more positive PQ-92 symptoms. To enhance the value of this sample for research purposes, six additional groups of approximately 25 individuals were randomly selected that endorsed 12, 13, 14, 15, 16 and 17 positive symptoms, respectively. Mean age was 25.7 (SD 5.0, range 16–35), and 31.4% were male (68.6% female).

3 *UHR follow-up*. A number of those individuals ($N = 90$) identified as being at UHR by the CAARMS in addition to a DSM-4 axis one (non-psychotic) or axis two diagnosis and that were willing to participate were followed up after 18 months (van der Gaag *et al.*, 2012). Mean age was 25.4 (SD 5.0, range 10–37), and 35.6% were male (64.4% female).

The two parameter logistic (2-PL) model and its assumptions

In this study we used the two parameter logistic (2-PL) model (Birnbaum, 1968), a type of IRT model appropriate to describe non-cognitive and clinical data (Reise and Waller, 2009). In the 2-PL model, the response probabilities of respondents to individual items are modeled by means of a logistic function whose precise form is defined by a discrimination and a location parameter. The discrimination parameter equals the slope of the logistic function and represents the discriminative power of the item (i.e. how much response probabilities are influenced by trait level). The location parameter equals the point of inflection (mean) of the logistic function and it also represents the intensity level of the item. These functions are also called item characteristic curves or item trace lines. In order to apply the 2-PL model, the related assumptions of unidimensionality and local independence must be met and the chosen model must fit the data reasonably well. Unidimensionality means that response behavior is influenced by one trait only, and local independence means that items are essentially uncorrelated when controlling for this trait. In IRT, positions of items and persons on the latent continuum are denoted as theta (θ). The distribution of persons on this latent continuum may be conceived as approximately standardized. The cutoff advised by Ising *et al.* (2012) for including patients in the CAARMS interview (more than 17 positive symptoms) corresponds with a θ -value of +0.81 on the positive symptom continuum (approximately highest 20%).

Model fit 2-PL and local dependence (LD)

To check IRT assumptions, we conducted the following analyses. To test for global fit, we compared the observed sum score distribution with the expected sum score distribution on the basis of the model. Large discrepancies indicate misfit. To check for local dependence (LD), we inspected the magnitudes of the residual correlations among the positive and negative symptoms respectively after fitting unidimensional models. We also checked item fit. The sample was divided into three groups of approximately equal size according to their score level (that is, total scores without the item targeted). These groups represent individuals with low, medium, and high levels of psychotic symptom experiences. Observed response probabilities within these groups were compared with model-based expected response probabilities, and mean absolute differences (MADs) were computed for each item. In this way, the appropriateness of the item trace line (logistic function) was evaluated for each item. All IRT-analyses were performed using the object-oriented, free available software package MIRT (Glas, 2010). The differences between observed and expected sum score frequencies and observed and expected response probabilities were evaluated using the Lagrange Multipliers (LM) test (Glas, 1999), which has an asymptotic chi-square distribution. In all applications of the LM test, absolute differences between observed and expected are more informative about model violations than the outcomes of the test statistics, as large sample sizes quickly lead to significant findings. The first sample (help-seeking) was used for these analyses.

Differential item functioning (DIF)

Because appropriateness of the item trace lines may also depend on the demographic background of respondents, it is important to investigate whether parameter estimates based on the whole sample are also appropriate

(invariant) for subgroups. Differential item functioning (DIF) tests essentially evaluate whether the increase in fit by freeing parameter estimates between groups is worth the number of additional parameters that have to be estimated. We investigated DIF for gender and age (adolescents versus adults). In order to check DIF for age, we split the data-file in to adolescents (<18 years; $N = 602$) and adults ($N = 5088$). We decided to consider MADs in response probabilities greater than 0.05 as moderate DIF and MADs greater than 0.10 as inadmissible for the purpose of adaptive testing (C. Glas, personal communication, February 6, 2015).

Simulation of CAT-properties based on item parameters and observed response patterns: RDS

RDS enable the evaluation of adaptive test properties before actually implementing the test. The estimated item parameters are used in combination with the observed response patterns to simulate an adaptive test (Sands *et al.*, 1997). The first item selected provides maximum information with regard to the group mean $(\theta_i = 0)$, and all subsequent items chosen for administration provide maximum information with regard to the estimated θ-values of each respondent. Based on the first sample, we (i) computed the correlation of θ -values obtained using the CAT-scores with full-length test-scores (θ-values based on the administration of all symptoms) and (ii) investigated measurement precision along the latent continua for CAT-POS and CAT-NEG. The program Firestar (Choi, 2009) was used to compile syntax to be used in *R* (R Core Team, 2014) to run these analyses. These simulated adaptive test scores were also used to investigate the criterion and predictive validity of the positive and negative symptom dimensions.

Criterion validity[1]

Combining structured interviews with indicators of patients' functioning is seen as the gold standard for the differentiation between healthy, UHR and psychotic individuals. In case of UHR, functioning must be either low, or recently declined in addition to the result of the interview. We used the CAARMS and the fourth version of the SOFAS for the differentiation between normal versus UHR/psychosis. The (Pearson) correlation, sensitivity, specificity, positive-predictive value (PPV), negative-predictive value (NPV) and the accuracy of the CAT-POS were compared with the same indices for the PQ-16. The second sample was used for these analyses. It has to be noted that the CAARMS-assessors were not blind to the PQ-scores of patients. That is, although assessors did not know precisely how many positive symptoms were endorsed by the patients they interviewed, they were sure that these patients endorsed at least 12 positive symptoms (inclusion criteria for the second sample).

Instruments

CAARMS

The CAARMS is a structured interview used to assess UHR status for psychosis. Reliabilities (intra-class correlations, ICC) for the positive symptom subscales that were used to define the UHR/psychosis status range from 0.79 to 0.89 for non-psychotic help-seeking individuals. The CAARMS discriminates well between healthy and UHR, and within UHR-samples, patients that are CAARMS positive are approximately 16 times more likely than CAARMS negative patients to develop a psychotic disorder (Yung *et al.*, 2008).

SOFAS

The SOFAS (Goldman *et al.*, 1992) assesses functioning on a scale ranging from 0 (poor functioning) to 100 (excellent functioning). Reliabilities (ICC or kappa) for the scale range from 0.55 to 0.80. SOFAS-scores have been consistently found to co-vary negatively with complexity of axis-one diagnosis and positively with other indicators of social and occupational functioning. Low functioning was operationalized as a score lower than 50, and substantial decline was operationalized as a drop of more than 30% from premorbid functioning (van der Gaag *et al.*, 2012).

Predictive validity

To explore the capability of the PQ-92-POS, CAT-POS, PQ-92-NEG, CAT-NEG and the PQ-16 of predicting important outcome criteria, a subgroup ($N = 90$) of the second sample was followed-up after 18 months. Outcome measures were the development of a psychotic disorder as diagnosed by psychiatrists, level of functioning measured by the SOFAS and the number of hospitalization days. The third sample (UHR follow-up) was used for these analyses and again (Pearson) correlations were computed. It should be noted that the third sample is not representative of help-seeking individuals because only patients classified as ARMS according to the CAARMS

were included. The attrition rate for this last stage of the data collection design was equal to 13%.

Results

Model fit 2-PL and local dependence

All IRT-analyses were conducted on the sample of help-seeking individuals.

Positive symptoms

Detailed output of the analyses is given in the Supportive Information; here we summarize the most important findings. Based on the LM test, we found significant differences between observed and expected sum score frequencies (LM = 80.14, $p < 0.01$). Closer inspection of these differences revealed that (i) especially zero scores are more frequently observed than the model implies and (ii) the differences are not systematic, in the sense that they do not show a clear pattern of deviation from the assumption of a normally distributed latent trait. The MADs between observed and expected response probabilities for the 45 symptoms were low, all between 0.00 and 0.01 with one of 0.02, meaning that the estimated item parameters fitted the observed responses quite well. Of the 990 item pairs $[(n * n - n)/2]$, nine had a residual correlation above 0.25, (maximum 0.34). The averaged absolute residual correlation was equal to 0.06, showing that the magnitudes of most correlations among positive symptoms were well reproduced by a unidimensional model.

Table 2 displays the SEs for 15 equally spaced intervals on the positive symptom continuum (all 45 items). SEs at the start of the continuum (very low scores) are higher than the SEs in the area surrounding the cutoff score for the CAARMS (θ = 0.81; 21.6% highest scores) or at the end of the latent continuum. This means that the 45 positive symptoms are less capable of differentiating among individuals who experience no or only a few mild symptoms than differentiating low scorers from those individuals that experience elevated levels of positive symptoms. Thus, we conclude that the positive symptom dimension may be adequately represented by the 2-PL model, noting that measurement precision is low at the beginning of the positive symptom continuum.

Negative symptoms

The differences between observed and expected sum score frequencies of the negative symptom dimension were not statistically significant (LM = 30.18, df = 19; not significant [n.s.]). However, as for the positive symptom dimension, the frequency of zero-scores is underestimated by the model. Again, the MADs between observed and expected response probabilities for the 19 symptoms were quite low (<0.01). Thus, negative symptoms can also be represented by the 2-PL model.

Of the 171 item pairs, three had a residual correlation above 0.25. The averaged absolute residual correlation was equal to 0.08 (maximum 0.44). Table 3 displays the SEs for 11 equally spaced intervals on the negative symptom continuum (all 19 items). For the negative symptom dimension, the differences in measurement precision along the latent continuum are smaller than was the case for the positive symptom dimension.

Differential item functioning (DIF)

Positive symptoms

On average men endorsed 0.6 positive symptoms less than women. Most positive symptoms displayed no DIF for gender and only one item displayed moderate DIF (MAD = 0.06): "I believe in telepathy, psychic forces, or fortune telling", with an LM-value of 109.2 (df = 1, sig. = 0.00). Men were a bit less (MAD = −0.08) likely to endorse this item than the model parameters suggested and women were somewhat more (+0.05) likely.

Adolescents endorsed 1.5 more positive symptoms than adults on average. Seven positive symptoms

Table 2. Number of respondents and averaged estimated standard errors (EAP) within 15 equally[1] spaced intervals (0.40) on the positive symptom dimension (all 45 symptoms)

	θ-intervals														
Min	−2.0	−1.6	−1.2	−0.80	−0.40	0.00	0.40	0.80	1.2	1.6	2.0	2.4	2.8	3.2	3.6
Max	−1.6	−1.2	−0.80	−0.40	0.00	0.40	0.80	1.2	1.6	2.0	2.4	2.8	3.2	3.6	3.8
N	289	245	366	417	501	510	507	412	285	243	135	94	56	22	4
SE	.55	.48	.43	.39	.35	.33	.30	.29	.27	.26	.26	.25	.26	.26	.21

[1]Min(θ_i) = −2.03, Max(θ_i) = 3.80.

Table 3. Number of respondents and averaged estimated standard errors (EAP) within 11 equally[1] spaced intervals (0.40) on the negative symptom dimension (all 19 symptoms)

	θ -intervals										
Min	−1.9	−1.4	−1.0	−0.60	−0.20	0.20	0.60	1.0	1.4	1.8	2.2
Max	−1.4	−1.0	−0.60	−0.20	0.20	0.60	1.0	1.4	1.8	2.2	2.5
N	385	317	428	535	666	629	556	413	256	131	71
SE	0.56	0.47	0.42	0.37	0.34	0.33	0.33	0.35	0.38	0.43	0.50

[1]Min(θ_i) = −1.84, Max(θ_i) = 2.49.

displayed moderate DIF for age, with MADs between 0.06 and 0.09. Detailed information on the results of these DIF tests can be found in Table A.3.0 in the Supporting Information.

Negative symptoms

On average, men endorsed 0.5 negative symptoms less than women, but all MADs where lower than 0.05. Adults endorsed 1.3 negative symptoms more than adolescents on average, but again, all MADs were lower than 0.05. In conclusion, the DIF-effects we found across subgroups were not substantial enough to justify the use of differential item parameters across groups.

Real data simulations (RDS)

Positive symptoms

Measurement precision of the positive symptom item pool was low for values lower than $\theta_i < -1.00$. Because we were not so much interested in how non-psychotic individuals score in terms of positive symptom experiences, but rather in differentiating between elevated and high levels, we used two stop criteria for these simulations: terminate the test session (i) if the upper bound of the 99.7% confidence interval $\left(\theta_i + 3^{*}\mathrm{SE}(\theta_i)\right)$ of the estimated score is lower than the corresponding cutoff score for CAARMS inclusion (21.6% highest scores) or (ii) when 12 items have been administered. A minimum of four items was always administered. When these boundary conditions were used, 10.1 items were utilized on average. The correlation of CAT-POS scores with full-length test scores (IRT-based) equaled 0.92 ($R^2 = 85\%$), indicating that both approaches yield roughly the same information. The average SE was equal to 0.47 ($r_{xx} = 0.82$), a value that is still slightly below the cutoff of 0.50 ($r_{xx} = 0.80$) (Evers *et al.*, 2010).

Negative symptoms

The following stop criteria were used: terminate the test sessions (i) if the corresponding SE is lower than 0.45 ($r_{xx} = 0.83$), or (ii) when 12 items have been administered. In this way, 8.8 items had to be utilized on average. The correlation with full-length test scores was 0.95 ($R^2 = 90\%$), and the SE equaled 0.46 ($r_{xx} = 0.83$) on average.

Criterion validity

The second sample is not representative for the target population of the screening tool (help-seeking population) because only individuals that endorsed many positive symptoms completed the CAARMS. In contrast to Ising *et al.* (2012), we chose not to impute CAARMS scores for the rest of the sample because (i) many ($N = 5132$) scores would have had to be imputed and (ii) we did not want to use positive symptoms as predictors for imputing CAARMS-scores (diffusion of predictor and criterion). Instead, we compared our results directly to those of Ising *et al.* (2012), using the same approach for the CAT-POS scores and the PQ-16. The correlations between the two predictors and the CAARMS were corrected for restriction of range in the predictor scores by Thorndike's case-2 formula (Wiberg and Sundström, 2009). The corrected correlation of the CAT-POS scores with the CAARMS (0.38) was lower than the correlation of the PQ-16 (0.47) with the CAARMS. Ising *et al.* (2012) advise to use a cutoff of more than 17 positive symptoms endorsed out of the 45 positive symptoms. Of the sample 21.6% endorse more than 17 positive symptoms, and this percentage corresponds to a θ-value above +0.81 on the CAT-POS. It should be noted that the goal was to differentiate between normal and UHR/psychosis, and not to identify individuals which are currently psychotic. The results for classification accuracy (healthy versus UHR/psychosis according to the CAARMS), using this value as cutoff, are displayed in Table 4.

Table 4. Classification accuracies of the PQ-16 and four different CAT-POS cutoff scores for CAARMS classifications

	Instrument & cutoff used				
	PQ-16	CAT-POS			
	$S_i > 5$	$\theta_i > 0.99$	$\theta_i > 0.94$	$\theta_i > 0.81$	$\theta_i > 0.62$
Sensitivity	0.93	0.74	0.75	*0.81*	0.84
Specificity	0.42	0.49	0.46	*0.39*	0.28
PPV	0.48	0.45	0.44	*0.44*	0.41
NPV	0.92	0.76	0.76	*0.78*	0.76
Accuracy	0.61	0.58	0.56	*0.54*	0.49

Note: cutoff advised by Ising *et al.* (2012) shown in italic typeface.

As shown in Table 4, the PQ-16 (second column) is superior to the CAT-POS (fifth column, $\theta_i > 0.81$) in terms of sensitivity (+12%), NPV (+14%) and accuracy (+7%). The differences in specificity and PPV are minimal. Because respondents in the second sample were selected by the number of positive symptoms they endorsed, it contains much less true negatives than might be expected without this restriction. Hence, the reported specificities, NPVs and accuracies underestimate the "true" values for both predictors. To a lesser degree, the reverse is also true for the reported sensitivities and PPVs of the CAT-POS and the PQ-16, because the selectiveness of the data collection design leads to a partial verification bias. Increasing the CAT-POS cutoff score (columns 3 and 4) would increase specificity and accuracy at the price of decreasing sensitivity. Increasing the CAT-POS cutoff score (columns 3 and 4) would increase specificity and accuracy at the price of decreasing sensitivity.

Predictive validity

Twenty-four (26.7%) patients transitioned to a psychotic state within the follow-up period of 18 month. The mean SOFAS score at the end of the follow-up period was equal to 55.70 (SD = 14.70). Eighty-one out of 90 patients (90%) were not hospitalized at all during the follow-up period, and the number of hospitalization days for the nine patients that did get hospitalized ranged from three to 230 days.

The correlations between the predictors and the follow-up criteria, calculated in the UHR follow-up sample, are displayed in Table 5. The CATs performed as well or even better than the unweighted symptom totals of the PQ-92. Although the PQ-16 was superior to the CAT-POS with respect to CAARMS classifications, the opposite was true for predicting (i) which patients will be diagnosed with psychotic disorder during the first 18 months after intake and (ii) the duration of hospitalization. No instrument predicted social and occupational functioning as assessed by the SOFAS, although the correlation for the CAT-NEG suggests an effect of interest, given the size of the correlation and the low p-value ($p = 0.058$).

Discussion

The present study showed the suitability of positive and negative symptoms of the PQ-92 for IRT-based adaptive

Table 5. Predictive validities of PQ-92-POS, CAT-POS, PQ-92-NEG, CAT-NEG and the PQ-16

	Diagnosis ($n = 90$)	Hospitalization days ($n = 89$)	SOFAS ($n = 78$)
PQ-92-POS	0.13	0.14	−0.08
CAT-POS	0.22*	0.24*	−0.06
PQ-92-NEG	0.10	0.05	−0.20
CAT-NEG	0.11	0.04	−0.20
PQ-16	0.14	0.17	−0.09

*$p < 0.05$ (one-tailed).

testing. Our results show that it is feasible to build CATs for these psychotic experiences, utilizing many fewer items than the original instrument, while yielding sufficient levels of measurement precision. On average, ten and nine items were required to place individuals on the positive and negative symptom dimensions, respectively. Although all effect sizes were small, in ARMS individuals (according to the CAARMS), the CAT-POS predicted best which individuals make a transition to psychosis and how long these would need to be hospitalized, and the CAT-NEG was suggestively associated with later functioning ($r_{xy} = -0.20$, $p = 0.058$; $N = 90$). With respect to CAARMS-classification accuracy, the PQ-16 was superior to the CAT-POS. It should be noted that during the RDS of the CATs, selection of those 16 items that make up the PQ-16 was rather the exception than the rule. With respect to the CAT-POS, symptoms of the facet Unusual Thought Content & Delusional Ideas were selected most frequently.

To our knowledge, this is the first study that applied IRT-models to the dimensions of positive and negative psychotic experiences. Because the symptoms were calibrated using a large sample drawn from the help-seeking population, the item parameters could be estimated precisely. Also, we used various different measures of validity to indicate how well the newly developed adaptive tests function. Several limitations should also be noted. The most important limitation is the selectness of the follow-up sample as a result of the data collection design: only individuals were included who had been classified as ARMS according to the CAARMS. Because of this, the results concerning the predictive validity of the CATs, while promising, should be replicated in a non-selected (including both ARMS and non-ARMS individuals) sample prospectively. This would enable a direct and fair comparison of the CATs on the one hand and the PQ-16 combined with the CAARMS on the other hand. With respect to the relationship between CAT-POS and number of hospitalization days, it is important to note that the distribution of hospitalization days was very skewed and the association that we found was based on only nine different time points. Although the RDS give an impression of how the CATs will function in practice and the results we found are promising, it has to be noted that the CATs were not yet applied in general practices. Furthermore, measurement precision for individuals that experience no or only a few positive or negative symptoms is lower than for individuals that experience elevated or high levels of symptoms. This finding is not uncommon for clinical scales, and the term *quasi-trait* has been introduced by Reise and Waller (2009) to describe this phenomenon: "(…) the trait is unipolar (relevant in only one direction) and that variation at the low end of the scale is less informative in both a substantive and psychometric sense." (p. 31). As such, neither item pool is ideal to track individual change (for example, when assessing recovery). Adding positive and negative symptoms with higher proportions of endorsement ($p > 0.70$) to the item pools would improve measurement precision at the low ends of the latent continua. It would be fruitful to investigate whether these indicators can be found in other inventories that assess milder forms of psychotic experiences.

The choice for CATs has two important advantages. First, in order to differentiate between positive and negative symptom experiences without the need of administering many items, the computed item parameters may be used in adaptive testing environments. We think that this benefit of economy is especially important in practical contexts where the aim is to assess a broad spectrum of psychopathological and psychological domains reliably without the need of administering hundreds of items in total, as is the case at the front door of the medical sector – that is, general practitioners' practices or cohort studies with a broad scope on diverse forms of psychopathology. In specialized secondary clinical settings where the focus is on specific psychopathological domains, diagnoses at intake are only preliminary and qualified CAARMS-assessors are available, this advantage will be probably less important, and thus the PQ-16 seems to be the better choice. Second, use of CATs offers the possibility of investigating the independent contribution of positive and negative symptom dimensions to the future development of psychosis and functional decline, without a priori capitalizing on present ARMS definitions according to the CAARMS. Within the current framework of ARMS (an important limitation of the present study), the CATs seems a promising and feasible concept to adequately and economically assess psychotic experiences, as CATs were associated with long-term outcomes (transition to psychosis, hospitalization and functioning). The item-parameters are provided in the Supporting Information and may be used for adaptive testing and computation of various IRT-metric scores.

Funding

This study was funded by a grant from the Mental Health Care Center Friesland, The Netherlands. J.T.W. Wigman was supported by Veni grant no 016.156.019. Data collection for the EDIE-NL study was supported by the

Netherlands Health Research Council, The Hague (ZonMW), 12051001; NTR1085 (awarded to Prof. Dr. Van der Gaag).

Declaration of interest statement

The authors have no competing interests.

Endnotes

1. We think that in this case, criterion validity is more appropriate a label than convergent validity, because, although CAARMS and CAT-POS assess the same domain, the methods are different (structured interview versus questionnaire). Furthermore, as long as the patients have not received a diagnosis by a psychiatrist, structured interviews are used as the gold standard for the differentiation between healthy, UHR and psychotic individuals.

References

Birnbaum A. (1968) Some latent trait models. In Lord F.M., Novick M.R. (eds) Statistical Theories of Mental Test Scores, pp. 397–479, MA, Addison-Wesley: Reading.

Chang W., Hui C., Tang J., Wong G., Chan S., Lee E., Chen E. (2013) Impacts of duration of untreated psychosis on cognition and negative symptoms in first-episode schizophrenia: a 3-year prospective follow-up study. *Psychological Medicine*, **43**(09), 1883–1893.

Chang W.C., Tang J.Y.M., Hui C.L.M., Lam M.M.L., Wong G.H.Y., Chan S.K.W., Chiu C.P.Y., Chung D.W.S., Law C.W., Tso S. (2012a) Duration of untreated psychosis: relationship with baseline characteristics and three-year outcome in first-episode psychosis. *Psychiatry Research*, **198**(3), 360–365.

Chang W.C., Tang J.Y., Hui C.L., Lam M.M., Chan S.K., Wong G.H., Chiu C.P., Chen E.Y. (2012b) Prediction of remission and recovery in young people presenting with first-episode psychosis in Hong Kong: a 3-year follow-up study. *The Australian and New Zealand Journal of Psychiatry*, **46**(2), 100–108.

Choi S.W. (2009) Firestar: computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement*, **33**(8), 644.

Embretson S.E., Reise S.P. (2013) Item Response Theory For Psychologists, Mahwah: Lawrence Erlbaum Associates.

Evers A., Lucassen W., Meijer R., Sijtsma K. (2010) COTAN Beoordelingssysteem voor de kwaliteit van tests, Amsterdam: Nederlands Instituut van Psychologen.

Glas C.A.W. (1999) Modification indices for the 2-PL and the nominal response model. *Psychometrika*, **64**(3), 273–294.

Glas C.A.W. (2010) Preliminary Manual of the Software Program Multidimensional Item Response Theory (MIRT). https://www.utwente.nl/bms/omd/medewerkers/temp_test/mirt-manual.pdf [6 January 2016].

Goldman H.H., Skodol A.E., Lave T.R. (1992) Revising axis V for DSM-IV: a review of measures of social functioning. *American Journal of Psychiatry*, **149**(9), 1148–1156.

Hanssen M.S.S. (2004) A Continuous Psychosis Phenotype: From Description To Prediction, PhD Series, Maastricht, South-Limburg Mental Health Research and Teaching Network

Ising H.K., Veling W., Loewy R.L., Rietveld M.W., Rietdijk J., Dragt S., Klaassen R.M., Nieman D.H., Wunderink L., Linszen D.H., van der Gaag M. (2012) The validity of the 16-item version of the Prodromal Questionnaire (PQ-16) to screen for ultra high risk of developing psychosis in the general help-seeking population. *Schizophrenia Bulletin*, **38**(6), 1288–1296.

Johns L.C., van Os J. (2001) The continuity of psychotic experiences in the general population. *Clinical Psychology Review*, **21**(8), 1125–1141.

Lin A., Wood S., Nelson B., Brewer W., Spiliotacopoulos D., Bruxner A., Broussard C., Pantelis C., Yung A. (2011) Neurocognitive predictors of functional outcome two to 13 years after identification as ultra-high risk for psychosis. *Schizophrenia Research*, **132**(1), 1–7.

Loewy R.L., Bearden C.E., Johnson J.K., Raine A., Cannon T.D. (2005) The prodromal questionnaire (PQ): preliminary validation of a self-report screening measure for prodromal and psychotic syndromes. *Schizophrenia Research*, **79**(1), 117–125.

Marshall M., Lewis S., Lockwood A., Drake R., Jones P., Croudace T. (2005) Association between duration of untreated psychosis and outcome in cohorts of first-episode patients: a systematic review. *Archives of General Psychiatry*, **62**(9), 975–983.

McGorry P.D., Killackey E., Yung A. (2008) Early intervention in psychosis: concepts, evidence and future directions. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, **7**(3), 148–156.

McGorry P.D., Yung A.R., Phillips L.J. (2003) The "close-in" or ultra high-risk model: a safe and effective strategy for research and clinical intervention in prepsychotic mental disorder. *Schizophrenia Bulletin*, **29**(4), 771–790.

Miller T.J., McGlashan T.H., Rosen J.L., Somjee L., Markovich P.J., Stein K., Woods S.W. (2002) Prospective diagnosis of the initial prodrome for schizophrenia based on the Structured Interview for Prodromal Syndromes: preliminary evidence of interrater reliability and predictive validity. *American Journal of Psychiatry*, **159**(5), 863–865.

Pogue-Geile M.F., Zubin J. (1987) Negative symptomatology and schizophrenia: a conceptual and empirical review. *International Journal of Mental Health*, **3**, 3–45.

Core Team R. (2014) R: A Language and Environment for Statistical Computing, Vienna: R Foundation for Statistical Computing.

Reise S.P., Waller N.G. (2009) Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, **5**, 27–48.

Sands W.A., Waters B.K., McBride J.R. (1997) Computerized Adaptive Testing: From Inquiry to Operation, DC, American Psychological Association: Washington.

Van der Gaag M., Nieman D.H., Rietdijk J., Dragt S., Ising H.K., Klaassen R.M., Koeter M., Cuijpers P., Wunderink L., Linszen D.H. (2012) Cognitive behavioral therapy for subjects at ultrahigh risk for developing psychosis: a randomized controlled clinical trial. *Schizophrenia Bulletin*, **38**(6), 1180–1188.

Van Os J., Hanssen M., Bijl R.V., Ravelli A. (2000) Strauss (1969) revisited: a psychosis continuum in the general population? *Schizophrenia Research*, **45**(1), 11–20.

Van Os J., Linscott R.J., Myin-Germeys I., Delespaul P., Krabbenda L. (2009) A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness–persistence–impairment model of psychotic disorder. *Psychological Medicine*, **39** (02), 179–195.

Van Os J., Verdoux H., Maurice-Tison S., Gay B., Liraud F., Salamon R., Bourgeois M. (1999) Self-reported psychosis-like symptoms and the continuum of psychosis. *Social Psychiatry and Psychiatric Epidemiology*, **34**(9), 459–463.

Wainer H. (2010) Computerized Adaptive Testing: A Primer, second edn, New York: Routledge.

Wiberg M., Sundström A. (2009) A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, **14**(5), 2.

Wigman J.T.W. (2011) Persistance of the Extended Psychosis Phenotype: Link between Vulnerability and Clinical Need, Ridderprint: Ridderkerk.

Wunderink L., Sytema S., Nienhuis F.J., Wiersma D. (2009) Clinical recovery in first-episode psychosis. *Schizophrenia Bulletin*, **35**(2), 362–369.

Yung A.R., McGorry P.D. (1996) The initial prodrome in psychosis: descriptive and qualitative aspects. *Australasian Psychiatry*, **30**(5), 587–599.

Yung A.R., Nelson B., Stanford C., Simmons M.B., Cosgrave E.M., Killackey E., Phillips L.J., Bechdolf A., Buckby J., McGorry P.D. (2008) Validation of "prodromal" criteria to detect individuals at ultra high risk of psychosis: 2 year follow-up. *Schizophrenia Research*, **105**(1), 10–17.

Yung A.R., Phillips L.J., McGorry P.D., McFarlane C.A., Francey S., Harrigan S., Patton G.C., Jackson H.J. (1998) Prediction of psychosis: a step towards indicated prevention of schizophrenia. *The British Journal of Psychiatry*, **172** (33), 14–20.

Yung A.R., Phillips L.J., Yuen H.P., Francey S.M., McFarlane C.A., Hallgren M., McGorry P.D. (2003) Psychosis prediction: 12-month follow up of a high-risk ("prodromal") group. *Schizophrenia Research*, **60**(1), 21–32.

Yung A.R., Yuen H.P., McGorry P.D., Phillips L.J., Kelly D., Dell'Olio M., Francey S.M., Cosgrave E.M., Killackey E., Stanfor C. (2005) Mapping the onset of psychosis: the Comprehensive Assessment of At-Risk Mental States. *Australian and New Zealand Journal of Psychiatry*, **39** (11–12), 964–971.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.